International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 9 Issue 5 ǁ 2021 ǁ PP. 57-63

What is Big Data and how Facebook uses Big Data

Remya.S.P
Assistant Professor On Contract,Department of Computer Science, University of Calicut,Kerala,India

Abstract: *With the advent and increased use of the internet, social media has become an integral part of people's daily routine. In the second quarter of 2020, Facebook has emerged as the biggest social network worldwide, with over 2.9 billion monthly active users which is approximately the 1/5th of the world's total population. Launched in the year 2004, it has grown tremendously since then. With a sudden obsession in social media, the number of people on Facebook has increased enormously, producing a massive amount of data every minute. They are generating almost 500 terabytes of data every day. To store and process this huge amount of data, the concept of big data is used. Big Data refers to large volume of data which is growing exponentially with time. Big Data uses open source data warehousing platforms like hadoop, hive etc. By being massively engaged in its capacity for collecting, storing, and interpreting data, Facebook has placed Big Data at the heart of its operations. This paper presents the application of big data on face book, examining its current as well as future impact.*

Keyword: *Big Data,Characteristics,Facebook,Types,Technology*

Date of Submission: 27-04-2021 Date of acceptance: 11-05-2021

I. INTRODUCTION

According to the current situation, we can strongly say that it is impossible to see a person without using social media. Because the world is getting drastic exponential growth digitally around every corner of the world. According to the report, from 2017 to 2021 the total number of social media users has been increased from 2.46 to 3.02 billion.

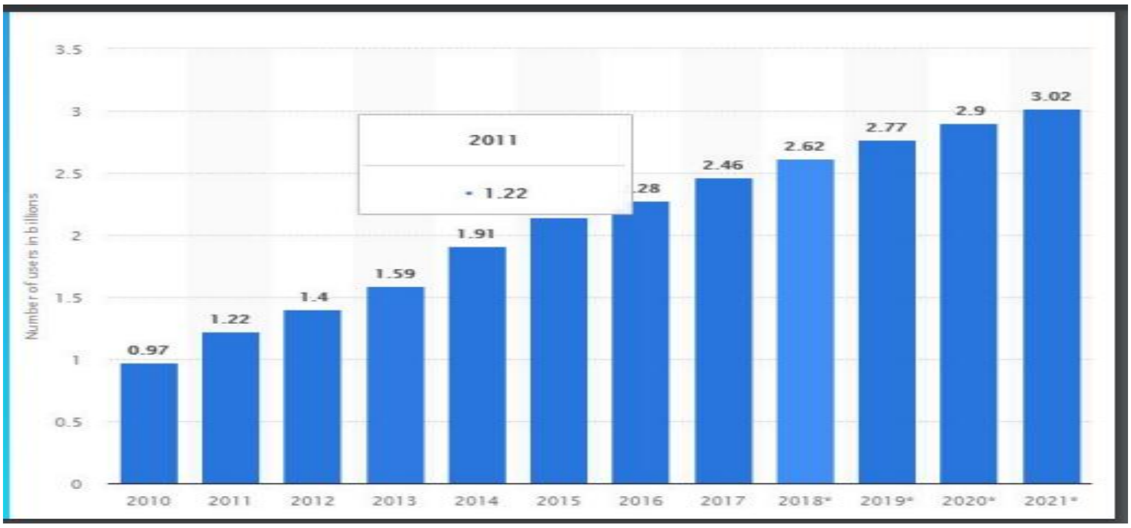


Fig.1. Graph showing increased number of social media users in each year

People are using Facebook, Youtube,Instagram, WhatsApp, and other social/Messaging medium while doing their daily routines. So, this caused the average time spent on social media by an individual has been increased to 2 hours 22 minutes**.**

Facebook stands today as one of the most popular social networking sites.The company stated that 3.14 billion people were using facebook, around 700,000 user's login to their account every minute, 136,000 photos are uploaded, 510,000 comments are posted, and 293,000 status updates are posted. They are generating almost 500 terabytes of data every day. That is a lot of data.

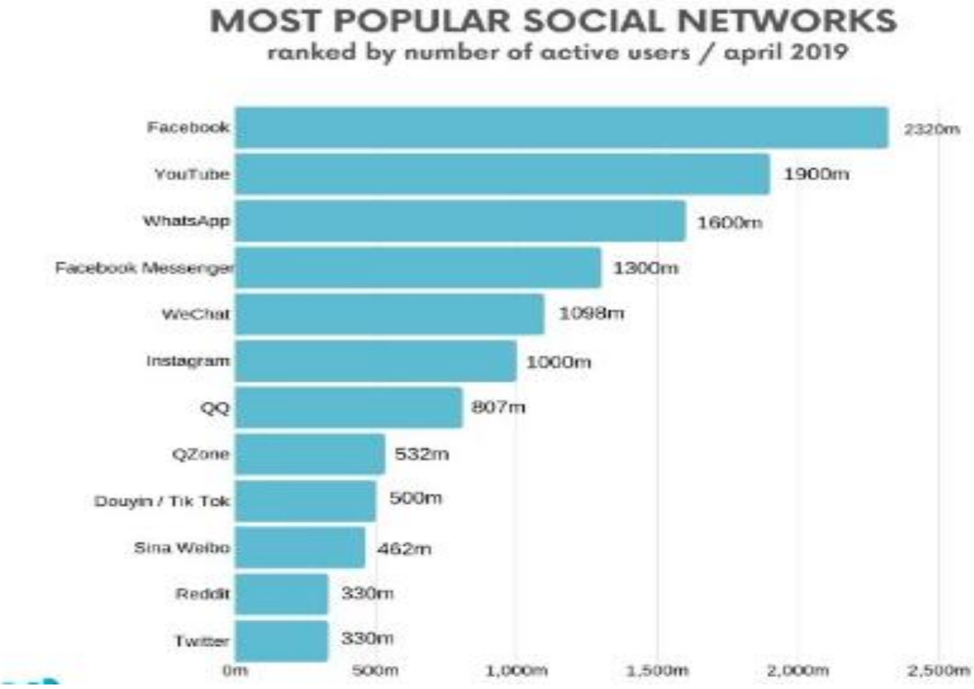


Fig.2. Most popular social networks ranked by number of active user's in April 2019

The amount of log and contextual data in facebook that needs to be processed and stored has exploded .At first, this information may not seem to mean very much. But with data like this, Facebook knows who our friends are, what we look like, where we are, what we are doing, our likes, our dislikes, and so much more.. To handle these huge volume of data, facebook has adopted BIG DATA technology. Big Data are generated continuously and are more flexible and scalable in their production. Big data offers vast opportunities whether used independently or with existing traditional data. Data scientists, analysts, researchers and business users can leverage these new data sources for advanced analytics that deliver deeper insights and to power innovative big data applications.

II. BIG DATA

2.1. What is Big Data?

Big Data is a collection of data that is huge in volume and size. The term big data has been used to describe data in the petabyte range or larger growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently.

2.2. Types of Big Data

Following are the types of Big Data:

A. Structured
B. Unstructured
C. Semi-structured

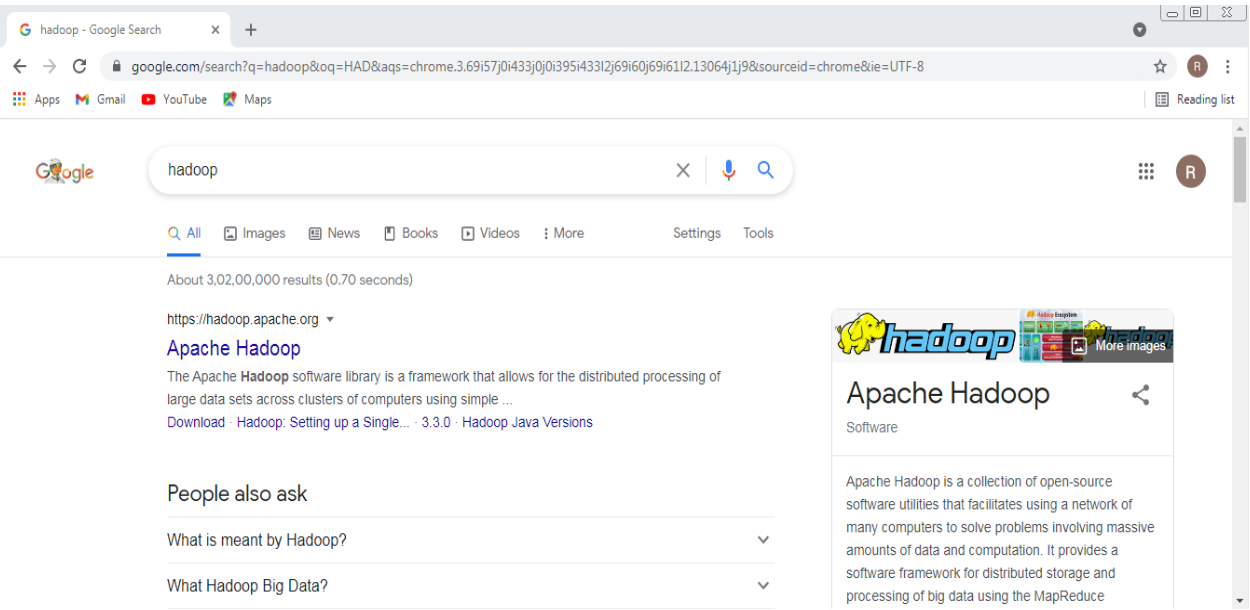
A. Structured data

They are the data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. Data stored in a relational database management system is one example of a 'structured' data.
An 'Employee' table in a database is an example of Structured Data

Employee_ID	Employee_Name	Gender	Department
101	Rohit	Male	IT
102	Ajay	Male	Finance
103	Deepika	Female	Admin

B. Unstructured data

Any data with unknown form or structure is classified as unstructured data. Example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.In addition to the size being large, un-structured data poses difficulty in processing these data . The output returned by 'Google Search' is an example Of Un-structured Data.



google.com/search?q=hadoop&oq=HAD&aqs=chrome.3.69i57j0i433j0j0i395i433l2j69i60j69i61l2.13064j1j9&sourceid=chrome&ie=UTF-8

hadoop

About 3,02,00,000 results (0.70 seconds)

https://hadoop.apache.org

Apache Hadoop

The Apache **Hadoop** software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple ...

Download · Hadoop: Setting up a Single... · 3.3.0 · Hadoop Java Versions

People also ask

What is meant by Hadoop?

What Hadoop Big Data?

Apache Hadoop

Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce

Fig.3. Example Of Un-structured Data

C. Semi-structured data

Semi-structured data can contain both the forms of data. It is structured in form but not defined. Example of semi-structured data is a data represented in an XML file. Examples of Semi-structured Data includes Personal data stored in an XML file-

```
<rec><name>Akash</name><sex>Male</sex><age>35</age></rec>
<rec><name>Radhika R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Shivkumar Mane</name><sex>Male</sex><age>29</age></rec>
```

2.3. Characteristics of Big Data

Big data can be described by the following characteristics:

A. Volume
B. Variety
C. Velocity

(A) Volume – The name Big Data itself is related to the size, whether a particular data is a Big Data or not, is dependent upon the volume of data. Thus Size of data plays a very crucial role in determining value out of data.

(B) Variety – Variety refers to the types or nature of data both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Now,data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

(C) Velocity – The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, mobile devices, etc. The flow of data is massive and continuous.

The convergence of social media and big data gives birth to a whole new level of technology.

III. APPLICATION OF BIG DATA IN FACEBOOK

Facebook is the world’s most popular social media network with more than two billion monthly active users worldwide. Facebook stores enormous amounts of user data, making it a massive data. Facebook is still under the top 100 public companies in the world, with a market value of approximately $475 billion. Every day, we feed Facebook’s data beast with mounds of information. Every 60 seconds, 136,000 photos are uploaded, 510,000 comments are posted, and 293,000 status updates are posted. The number of people on Facebook has increased enormously, producing a massive amount of data every minute. Every time you open your Facebook account, on an average you notice more than 10 new feed items, 5 notifications, 10 messages, and two friend requests. That is a huge amount of big data. The main business strategy of Facebook is to understand who their user's are by understanding their behaviour interest and their geographic location Facebook shows customized ads on the user's timeline. Apart from Google, Facebook is probably the only company that possesses this high level of detailed customer information. How it is possible? There are around billion levels of unstructured data has been generated everyday which contains images, text, videos and everything .With the help of Deep Learning Methodology (Artificial Intelligence), Facebook brings structure for large volume of unstructured data.

Here are a few examples that show how Facebook uses its Big Data.

Example 1: The Flashback videos

In honor of its 10th anniversary, Facebook offered its users the option to view and share a video that traces the course of their social network activity from the date of registration until the present. This one minute video, called the "Flashback," is a collection of photos and posts that received the most "likes" or comments, all set to nostalgic background music. These videos are also presented to users on occasions like their "friendversary", i.e. the anniversary of the day they became friends on the platform or on the occasion of the user's birthday. This mini-movie symbolizes how Facebook can make of our data.

Example 2: People you may know

One of the most controversial parts of facebook data collection is a feature called "people you may know" .This is where facebook uses many different signals of what it knows about you to determine who else you might be connected to. And this is not always things that you share with facebook, it might be contacts in your phone or it users location data to recommend friends.

Example 3:- Facial recognition

One of Facebook's latest investments has been in facial recognition and image processing capabilities– a technology that instructs machines on how to detect the details in a specific picture or video, by guiding it through various other images. When you upload a picture, you might see suggestions for people you could tag on it. This is based on analysis of the picture data, which is compared against pictures of people in your Friends list. Facebook can track its users across the internet and other Facebook profiles with image data provided through user sharing. The Deep Learning application **"DeepFace"** is the tool used to detect people in pictures. The platform claims that its most advanced image recognition tool is more successful than humans in detecting if two different images are of the same person or not.

Example 4: Tracking cookies

Facebook tracks its users across the web by using tracking cookies. If a user is logged into Facebook and simultaneously browses other websites, Facebook can track the sites they are visiting..

Example 5: Target Advertisements

Facebook uses deep neural networks to decide how to target audience while advertising ads. Big data strategy and predictive analytics make it possible to improve decision making on the basis of past history. Data-driven business tends to succeed enormously as computers can provide forthcoming customer choices. Though interests and habits change with time, in general, they remain related. Once a user buys something, there is a great possibility of choosing similar products. Because of this serving, the highly targeted advertising, Facebook has become the toughest competitive for the ever known search engine Google.

IV. TECHNOLOGY USED BEHIND FACEBOOK'S BIG DATA

There is a combined workforce of people and technology constantly working behind the successful implementation of this platform. Facebook Inc. analytics chief Ken Rudin says, "Big Data is crucial to the company's very being." He goes on to say that, "Facebook relies on a massive installation of Hadoop, a highly scalable open-source framework that uses clusters of low-cost servers to solve problems. Facebook even designs its hardware for this purpose. Hadoop is just one of many Big Data technologies employed at Facebook."

Though the platform is continuously being enriched, below are the prime technological aspects:

(A) Hadoop

"Facebook runs the world's largest Hadoop cluster" says Jay Parikh, Vice President Infrastructure Engineering, Facebook.

Basically, Facebook runs the biggest Hadoop cluster that goes beyond 4,000 machines and storing more than hundreds of millions of gigabytes. This extensive cluster provides some key abilities to developers:

- The developers can freely write map-reduce programs in any language.
- SQL has been integrated to process extensive data sets, as most of the data in Hadoop's file system are in table format. Hence, it becomes easily accessible to the developers with small subsets of SQL.

Hadoop provides a common infrastructure for Facebook with efficiency and reliability. Beginning with searching, log processing, recommendation system, and data warehousing, to video and image analysis, Hadoop is empowering this social networking platform in each and every way possible. Facebook developed its first user-facing application, Facebook Messenger, based on Hadoop database, i.e., Apache HBase, which has a layered architecture that supports plethora of messages in a single day.

(B) Scuba

With a huge amount of unstructured data coming across each day, Facebook slowly realized that it needs a platform to speed up the entire analysis part. That's when it developed **Scuba**, which could help the Hadoop developers dive into the massive data sets and carry on ad-hoc analyses in real-time. Facebook was not initially prepared to run across multiple data centers and a single break-down could cause the entire platform to crash. Scuba, another Big data platform, allows the developers to store bulk in-memory data, which speeds up the informational analysis. It implements small software agents that collect the data from multiple data centers and compresses it into the log data format. Now this compressed log data gets compressed by Scuba into the memory systems which are instantly accessible. According to Jay Parikh, "Scuba gives us this very dynamic view into how our infrastructure is doing — how our servers are doing, how our network is doing, how the different software systems are interacting."

(C) Cassandra

"The amount of data to be stored, the rate of growth of the data, and the requirement to serve it within strict SLAs made it very apparent that a new storage solution was absolutely essential."**-** Avinash Lakshman , Search Team, and Facebook. The traditional data storage started lagging behind when Facebook's search team discovered an Inbox Search problem. The developers were facing issues in storing the reverse indices of messages sent and received by the users. The challenge was to develop a new storage solution that could solve the Inbox Search Problem and similar problems in the future. That is when Prashant Malik and Avinash Lakshman started developing **Cassandra**. The objective was to develop a distributed storage system dedicated to managing a large amount of structured data across multiple commodity servers without failing once.

(D) Hive

After Yahoo implemented Hadoop for its search engine, Facebook thought about empowering the data scientists so that they could store a larger amount of data in the Oracle data warehouse. Hence, Hive came into existence. This tool improved the query capability of Hadoop by using a subset of SQL and soon gained popularity in the unstructured world. Today almost thousands of jobs are run using this system to process a range of applications quickly.

(E) Prism

Hadoop wasn't designed to run across multiple facilities. Typically, because it requires such heavy communication between servers, clusters are limited to a single data center.Initially when Facebook implemented Hadoop, it was not designed to run across multiple data centers. And that's when the requirement to develop Prism was felt by the team of Facebook. Prism is a platform which brings out many namespaces instead of the single one governed by the Hadoop. This in turn helps to develop many logical clusters.This system is now expandable to as many servers as possible without worrying about increasing the number of data centers.

(F) Corona

Developed by an ex-Yahoo man Avery Ching and his team, Corona allows multiple jobs to be processed at a time on a single Hadoop cluster without crashing the system. This concept of Corona sprouted in the minds of developers, when they started facing issues with Hadoop's framework. It was getting tougher to manage the cluster resources and task trackers. Map Reduce was designed on the basis of a pull-based scheduling model, which was causing a delay in processing the small jobs. Hadoop was limited by its slot-based resource management model, which was wasting the slots each time the cluster size could not fit the configuration. Developing and implementing Corona helped in forming a new scheduling framework that could separate the cluster resource management from job coordination.

(G) Peregrine

Another technological tool that is developed by Murthy was Peregrine, which is dedicated to addressing the issues of querying data as quickly as possible. Since Hadoop was developed as a batch system that used to take time in running different jobs, Peregrine brought the entire process close to real-time.

Apart from the above prime implementations, Facebook uses many other small and big sized pieces of technology to support its Big Data infrastructure, such as Memcached, Hiphop for PHP, Haystack, Bigpipe, Scribe, Thrift, etc.

Today Facebook is one of the biggest corporations on earth thanks to its extensive data on over one and a half billion people on earth. This has given it enough clout to negotiate with over 3 million advertisers on its platform in order to clock staggering revenues that is north of 17 Billion US Dollars. But the privacy and security concerns still loom large regarding whether Facebook will utilize all that gargantuan volumes of data to server humanity's greater good or just use it to make more money.

V. DRAWBACKS OF ADOPTION OF BIG DATA

A. Raises privacy concerns

The adoption of big data by Facebook has raised many privacy issues .Many users complain that the privacy settings of the platform are complicated or not properly clarified which results in people sharing information which they didn't intend to. While the platform has repeatedly attempted to adjust this, this has ended up with people getting muddled since many of them had grown accustomed to the existing features of the platform.

B. Issues encountered over face recognition

The platform's facial recognition tool has raised a plethora of doubts and concerns. In this tool, once a picture is uploaded, the user gets suggestions regarding the people they can tag in it. This tool is based on an interpretation of the picture's data, which is compared against pictures of the people in the user's Friends list. The adoption of this technology has proven to be pretty controversial. Various privacy campaigners claim that the application exceeds its limits since, in the case of high-resolution pictures of a crowd, it permits Facebook to identify a lot of the faces which becomes a hindrance to the public's privacy and freedom to be anonymous. Recently the platform had consented to pay $650 million or a long-lasting class-action lawsuit, regarding its adoption of facial recognition. This lawsuit is regarding the platform's photo-tagging feature, for which it employs facial recognition software for identifying faces in the user's photos. The state of Illinois has established a law opposing the businesses gathering biometric data without obtaining consent firsthand and the state claimed that Facebook didn't obtain approval prior to allowing the new feature by default to all the Facebook users in the state owing to which it sued the company in 2015.

C. Big Data Analytics At Facebook

The manner in which the habits of the user are assessed has raised additional concerns. In January 2020, Facebook introduced its Off Facebook Activity tracker. The tool offers users an itemized list of the websites, apps, and real-life stores that the platform knows they visited and also gives them the option to disable the tracking. Through these latest assessing tools the platform records everything the user does from how long their cursor hovers over certain parts of a page to what websites they visit outside of the platform, which in turn is used for generating the appropriate algorithms to determine the kinds of adverts to show the users. Facebook has its own Data Science team with its own page where they consistently post updates regarding the insights which they have gathered from assessing the habits of the million people browsing the site. From predicting the intelligence of users, their perspective on political issues, or their emotional stability, big data has played an integral role in making the platform effectively understand their users. This is an effort to aid the platform in selling targeted advertisements, yet it, in turn, raises the dilemma of how the data can be used by the government for discovering the public views and manipulating them.

VI. CONCLUSION

There is no doubt that Facebook is one of the largest Big Data specialists, dealing with petabytes of data, including historical and real-time, and will keep growing in the same horizon. While the world is coming closer together on this platform, Facebook uses technologies to track those connections and their presence on or outside its walls to fetch the most suitable posts for its users. From our friends, our location, our looks, our occupation to our likes and dislikes, Facebook has a lookout for everything! The platform possesses a detailed and advanced level of data regarding its users. The more users that adopt Facebook, the more data is garnered by the platform. By being massively engaged in its capacity for collecting, storing, and interpreting data, Facebook has placed Big Data at the heart of its operations. One thing is for sure, it is Big Data that has propelled Facebook, a small-time Harvard dorm startup into the constellation of some of the biggest corporations on earth of all times!

REFERENCES

[1] Steve Lohr,. "The Origins of 'Big Data': An Etymological Detective Story". The New York Times. 28 September 2016.

[2] Oracle and FSN, "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity" , December 2012

[3] Jacobs, A,"The Pathologies of Big Data". ACMQueue, 6 July 2009.

[4] Sagiroglu, Seref, "Big data: A review".,2013 International Conference on Collaboration Technologies and Systems ISBN 978-1-4673-6404-1. .

[5] Kitchin, Rob; McArdle, Gavin "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets". 17 February 2016.

[6] "Big Data",https://en.wikipedia.org

[7] Data, data everywhere. The Economist (http://www.economist.com/node/15557443), Feb 2010.

[8] IBM " What is big data? – Bringing big data to the enterprise". ibm.com.

[9] V Marx "The big challenges of big data"- Nature, 2013 - nature.com

[10] M Chen, S Mao, Y Liu "Big data: A survey on Mobile networks and applications", 2014 - Springer

[11] Francis J. Alexander, ,Adolfy Hoisie, ,Alexander Szalay, "Computing in Science & Engineering" Nov.-Dec. 2011, vol. 13

[12] Alexandros Labrinidis,H. V. Jagadish "Challenges and opportunities with big data"

[13] Y. Noguchi "The Search for Analysts to Make Sense of Big Data." National Public Radio

[14] S. Lohr, "The age of big data", New York Times (http://www.nytimes.com), Feb 2012

[15] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers,"Big data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute, May 2011

[16] Elena Kvochko, "Four Ways To talk About Big Data (Information Communication Technologies for Development Series)". worldbank.org. 4 December 2012.

[17] Bernhard Warner "'Big Data' Researchers Turn to Google to Beat the Markets". Bloomberg Businessweek, 25 April 2013.

[18] Viktor Mayer-Schönberger; Kenneth Cukier ," Big Data: A Revolution that Will Transform how We Live, Work, and Think". Houghton Mifflin Harcourt, 2013,ISBN 9781299903029

[19] Press, Gil ,"A Very Short History of Big Data". forbes.com. Jersey City, NJ: Forbes Magazine,17 September 2016

[20] Darwin Bond-Graham, *The Perspective on Big Data,* ThePerspective.com, 2018